ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Clasificación de pacientes escolares con posible trastorno de déficit de atención e hiperactividad (TDAH) mediante técnicas de Machine Learning

PROYECTO INTEGRADOR

Previo a la obtención del Título de:

Matemático

Presentado por:

Luis Diego Campuzano Almeida

Hopkins Agustin Herbas Chaverro

GUAYAQUIL - ECUADOR

Año: 2023

DEDICATORIA

A mis padres que en paz descansen, a mí papá Luis Campuzano quien supo darme los consejos oportunos para cada etapa de mi vida, y a mi madre Bethzabe Almeida, que inculcó en mí los estudios, la superación, la perseverancia, y la responsabilidad para llegar al lugar en que estoy en estos momentos. Este trabajo está dedicado a ustedes con mucho amor. Su hijo Luis Campuzano.

Luis Campuzano A.

DEDICATORIA

Para los que día a día sufren y soportan las injusticias de este mundo. A los inocentes y los que trabajan día y noche.

Esta tesis es para ustedes.

Hopkins Herbas C.

AGRADECIMIENTOS

Sin lugar a dudas tengo que dar gracias principalmente a Jehová Dios, ya que gracias a su misericordia llegué al lugar donde estoy ahora, y es por él que las cosas se dieron. Quiero hacer una mención especial a mi primo Alvaro Criollo que, en un momento oscuro de mi vida, me sacó del hoyo en el que me encontraba, él fue mi guía y mi apoyo, quien me encaminó a retomar mis estudios.

También doy las gracias a todas esas personas que me dieron un plato de comida, que me dieron una prenda de vestir, que me cuidaron cuando estaba enfermo, aquellas que me brindaron su compañía. La lista de esas personas es inmensa, pero para aquellas que lean esto les doy las gracias.

Gracias a mis profesores, que a su debido tiempo supieron darme consejo, me escucharon, me motivaron, me educaron, y me llenaron de confianza para seguir adelante. Le doy gracias a Dios por haberlos conocidos y de que sean mis profesores.

Luis Campuzano A.

AGRADECIMIENTOS

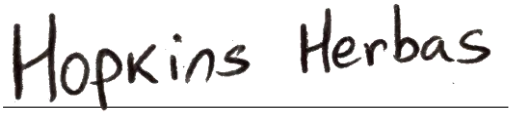
Primeramente agradezco al único que siempre estuvo ahí presente, a Dios. Agradezco a mi familia por enseñarme que la vida muchas veces es injusta y que a pesar de ello los sueños se cumplen cuando se trabaja por ellos. A mis mascotas las cuales no hablan pero sus miradas lo dicen todo. A todas esas buenas personas que me ayudaron en el momento exacto, desde un almuerzo hasta una laptop, porque esas son las acciones que han cambiado mi vida. A todos los amigos y compañeros que hice mientras cursaba mi carrera. A los profesores, por compartirme sus conocimientos y agradezco mucho la gran paciencia que tienen. A Cristiano Ronaldo, por enseñarme a ser mejor cada día.

Hopkins Herbas C.

DECLARACIÓN EXPRESA

“Los derechos de titularidad y explotación, nos corresponde conforme al reglamento de propiedad intelectual de la institución; *Luis Diego Campuzano Almeida, Hopkins Agustin Herbas Chaverro,* y damos nuestro consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual”

Luis Campuzano A.



Hopkins Herbas C.

EVALUADORES

Firmado electrónicamente por:
LUZ ELIMAR MARCHAN MENDOZA

Luz Elimar Marchán Mendoza

PROFESOR DE LA MATERIA

Firmado electrónicamente por:
HEYDI MARIANA ROA LOPEZ

Heydi Mariana Roa López

TUTOR

RESUMEN

El trastorno por déficit de atención e hiperactividad (TDAH) es un trastorno neurobiológico que se caracteriza por problemas de atención, hiperactividad e impulsividad. Se ha convertido en uno de los trastornos más comunes en la infancia, que afecta mayoritariamente a niños de hasta 12 años y en muchos casos persiste hasta la edad adulta. El TDAH afecta significativamente el desempeño académico y laboral del individuo, así como también en sus relaciones interpersonales y autoestima. En el Ecuador, son aún muy pocos los estudios sobre esta patología. Sin embargo, los estudios a nivel mundial indican que un diagnóstico temprano y la aplicación de un tratamiento adecuado permiten llevar a estas personas, una vida plena y satisfactoria. Por ello, la importancia del diagnóstico temprano de los infantes por parte del profesional de la salud mental. Pero evaluar a todos los niños con las herramientas dispuestas por los profesionales conlleva un reto enorme.

En este trabajo se propone predecir el diagnóstico de un infante mediante algoritmos de Machine Learning. Se implementaron tres modelos de clasificación K- medias, K vecinos más cercanos, y el que resultó más eficiente, Máquinas de Vectores de Soporte (SVM). Se usaron los datos de las respuestas a dos tests psicológicos (Conners y Brief) de niños de segundo, cuarto y sexto de educación básica, que incluían también las de sus respectivos padres y profesores.

Los resultados obtenidos muestran que el ordenador puede predecir si un niño posee o no TDAH con un 85% de precisión. Finalmente, este trabajo tiene como propósito servir a los expertos en el área, como una herramienta que ofrece una reducción importante de tiempo en la identificación del trastorno en los niños, además de ahorro económico.

Palabras Clave: TDAH, clasificación, máquinas de vectores de soporte, Test psicológico.

ABSTRACT

Attention deficit hyperactivity disorder (ADHD) is a disease that mainly affects children up to 12 years of age, and directly alters the academic performance of the child. It has been little explored in Ecuador. The process of diagnosis is an essential phase for early treatment for children, but evaluating all children with the tools provided by psychologists is a huge challenge.

We propose to predict the diagnosis of an infant using Machine Learning algorithms. In this work we implemented 3 classification models K-means, K nearest neighbors and the most important one, Support Vector Machines (SVM). The database used were 2 psychological tests (Conners and Brief) whose content included the results of the same, and the respondents were children of second, fourth and sixth grade of basic education, as well as their respective parents and teachers.

The results obtained show that the computer can predict whether or not a child has ADHD with 85% accuracy. Finally, this work can serve as a time- and cost-saving tool for experts in the field to identify the disease in children.

Keywords: *ADHD, classification, support vector machines, Psychological test.*

ÍNDICE GENERAL

RESUMEN I
ABSTRACT II
ABREVIATURAS V
SIMBOLOGÍA VI
ÍNDICE DE FIGURAS VII
ÍNDICE DE TABLAS IX
CAPÍTULO 1 1
1. INTRODUCCIÓN 1
1.1 Descripción del problema 1
1.2 Justificación del problema 2
1.3 Objetivos 3
1.3.1 Objetivo General 3
1.3.2 Objetivos Específicos 3
1.4 Marco teórico 4
1.5 Definiciones claves 6
1.6 Tests para diagnosticar TDAH 6
1.7 Algoritmos de Machine Learning para la clasificación 7
1.7.1 K- Medias de agrupamiento 7
1.7.2 1.7.2 K- vecinos más próximos (K- nn) 8
1.7.3 Máquinas de Vectores de Soporte (SVM) 10
CAPÍTULO 2 12
2. METODOLOGÍA 12
2.1 Descripción de la muestra 12
2.2 Preparación de los datos 12
2.3 Modelos de Machine Learning 13
2.4 Elección de variables 14
2.5 Implementación del Modelo K-Means 16

2.6 Implementación del Modelo SVM 17
2.7 Implementación del Modelo K-NN 17
2.8 Análisis del porcentaje de error de cada modelo 18
2.9 Gráficas del Modelo SVM 18
2.10 Matriz de Confusión 18
CAPÍTULO 3 20
3. RESULTADOS Y ANÁLISIS 20
3.1 Análisis de resultados 20
3.2 Máquina de Vectores de Soporte 20
3.3 Porcentaje de eficacia 22
3.4 Matriz de Confusión 23
3.5 Gráficos de modelo SVM 24
CAPÍTULO 4 26
4. CONCLUSIONES Y RECOMENDACIONES 26
BIBLIOGRAFÍA . . . . . .
APÉNDICES
!
Apéndice A . . . . . .
Apéndice B . . . . . .
Apéndice C . . . . . .

ABREVIATURAS

ESPOL	Escuela Superior Politécnica del Litoral
SVM	Support Vector Machines
TDAH	Trastorno de déficit de atención e hiperactividad
ML	Machine Learning
KNN	K Nearest Neighbor
SHC	Estudiantes Hombres de segundo grado evaluados con el Test de Conners
SMC	Estudiantes Mujeres de segundo grado evaluados con el Test de Conners
CHC	Estudiantes Hombres de cuarto grado evaluados con el Test de Conners
CMC	Estudiantes Mujeres de cuarto grado evaluados con el Test de Conners
XHC	Estudiantes Hombres de sexto grado evaluados con el Test de Conners
XMC	Estudiantes Mujeres de sexto grado evaluados con el Test de Conners
SHB	Estudiantes Hombres de segundo grado evaluados con el Test de Brief
SMB	Estudiantes Mujeres de segundo grado evaluados con el Test de Brief
CHB	Estudiantes Hombres de cuarto grado evaluados con el Test de Brief
CMB	Estudiantes Mujeres de cuarto grado evaluados con el Test de Brief
XHB	Estudiantes Hombres de sexto grado evaluados con el Test de Brief
XMB	Estudiantes Mujeres de sexto grado evaluados con el Test de Brief

SIMBOLOGÍA

$N(u)$	Cardinalidad del conjunto U.
$\bigcup_{i=1}^{K} S_i$	Unión numerable de K conjuntos.
$d(X,Y)$	Distancia euclideana de X a Y.
$S_i^{(t)}$	Conjunto S_i en la iteración número t.
$\mathbb{R}^n$	Conjunto de los vectores x pertenecientes a $\mathbb{R}^n$
$\|\cdot\|$	Norma-p vectorial

ÍNDICE DE FIGURAS

Figura 1.1 Se busca asignar una etiqueta a la muestra (círculo gris), en este caso K=7, se toman los siete vecinos más cercanos, de los cuales el grupo de los cuadrados posee mayor número de representantes (3) por lo que la muestra se asigna a dicho grupo. 9
Figura 1.2 Interpretación gráfica del método del Kernel 10

Figura 2.1 Matriz de Confusión para los conjuntos de datos con el cuestionario de Conners 19
Figura 2.2 Matriz de Confusión para los conjuntos de datos con el cuestionario de Brief 19

Figura 3.1 Gráfico del modelo SVM grupo de Segundo Hombres test de Conners . 24

Figura B.1 Gráfico del modelo SVM grupo de Segundo Mujeres test de Conners. .
Figura B.2 Gráfico del modelo SVM grupo de Cuarto Hombres test de Conners. . .
Figura B.3 Gráfico del modelo SVM grupo de sexto Hombres test de Conners. . .
Figura B.4 Gráfico del modelo SVM grupo de cuarto Mujeres test de Conners. . . .
Figura B.5 Gráfico del modelo SVM grupo de sexto mujeres test de Conners. . . .
Figura B.6 Gráfico del modelo SVM grupo de Segundo Hombres test de Brief. . . .
Figura B.7 Gráfico del modelo SVM grupo de Segundo Mujeres test de Brief. . . .
Figura B.8 Gráfico del modelo SVM grupo de cuarto hombres test de Brief. . . . .
Figura B.9 Gráfico del modelo SVM grupo de Cuarto mujeres test de Brief. . . . .
Figura B.10 Gráfico del modelo SVM grupo de Sexto Hombres test de Brief. . . . .
Figura B.11 Gráfico del modelo SVM grupo de sexto mujeres test de Brief. . . . . .

ÍNDICE DE TABLAS

Tabla 2.1 Variables Test de Conners 15

Tabla 3.1 Vectores Soportes pertenecientes a la prueba de Conners. 21
Tabla 3.2 Vectores Soportes pertenecientes a la prueba de Brief. 21
Tabla 3.3 Porcentajes de error de los 12 grupos de estudio tanto para la data de entrenamiento como para la data de prueba. 22
Tabla 3.4 Matriz de confusión perteneciente al grupo de Segundo Hombres test de Conners, para los datos de entrenamiento. 23
Tabla 3.5 Matriz de confusión perteneciente al grupo de Segundo Hombres test de Conners, para los datos de Prueba. 24

Tabla A.1 Matriz de confusión perteneciente al grupo de Segundo Mujeres test de Conners, para los datos de entrenamiento. . . . . .
Tabla A.2 Matriz de confusión perteneciente al grupo de Segundo Mujeres test de Conners, para los datos de Prueba. . . . . .
Tabla A.3 Matriz de confusión perteneciente al grupo de Cuarto Hombres test de Conners, para los datos de Entrenamiento. . . . . .
Tabla A.4 Matriz de confusión perteneciente al grupo de Cuarto Hombres test de Conners, para los datos de Prueba. . . . . .
Tabla A.5 Matriz de confusión perteneciente al grupo de Cuarto Mujeres test de Conners, para los datos de entrenamiento. . . . . .
Tabla A.6 Matriz de confusión perteneciente al grupo de Cuarto Mujeres test de Conners, para los datos de Prueba. . . . . .
Tabla A.7 Matriz de confusión perteneciente al grupo de Sexto Hombres test de Conners, para los datos de Entrenamiento. . . . . .
Tabla A.8 Matriz de confusión perteneciente al grupo de Sexto Hombres test de Conners, para los datos de Prueba. . . . . .

Tabla A.9 Matriz de confusión perteneciente al grupo de Sexto Mujeres test de Conners, para los datos de Entrenamiento. . . . . .
Tabla A.10 Matriz de confusión perteneciente al grupo de Sexto Mujeres test de Conners, para los datos de Prueba. . . . . .
Tabla A.11 Matriz de confusión perteneciente al grupo de Segundo Hombres test de Brief, para los datos de entrenamiento. . . . . .
Tabla A.12 Matriz de confusión perteneciente al grupo de Segundo Hombres test de Brief, para los datos de prueba. . . . . .
Tabla A.13 Matriz de confusión perteneciente al grupo de Segundo Mujeres test de Brief, para los datos de entrenamiento. . . . . .
Tabla A.14 Matriz de confusión perteneciente al grupo de Segundo Mujeres test de Brief, para los datos de prueba. . . . . .
Tabla A.15 Matriz de confusión perteneciente al grupo de Cuarto Hombres test de Brief, para los datos de entrenamiento. . . . . .
Tabla A.16 Matriz de confusión perteneciente al grupo de Cuarto Hombres test de Brief, para los datos de prueba. . . . . .
Tabla A.17 Matriz de confusión perteneciente al grupo de Cuarto Mujeres test de Brief, para los datos de entrenamiento. . . . . .
Tabla A.18 Matriz de confusión perteneciente al grupo de Cuarto Mujeres test de Brief, para los datos de prueba. . . . . .
Tabla A.19 Matriz de confusión perteneciente al grupo de Sexto Hombres test de Brief, para los datos de entrenamiento. . . . . .
Tabla A.20 Matriz de confusión perteneciente al grupo de Sexto Hombres test de Brief, para los datos de prueba. . . . . .
Tabla A.21 Matriz de confusión perteneciente al grupo de Sexto Mujeres test de Brief, para los datos de entrenamiento. . . . . .
Tabla A.22 Matriz de confusión perteneciente al grupo de Sexto Mujeres test de Brief, para los datos de prueba. . . . . .

Tabla C.1 Variables Test de Brief . . . . . .
Tabla C.2 Variables Test de BRIEF . . . . . .

CAPÍTULO 1

1. INTRODUCCIÓN

El trastorno de déficit de atención e hiperactividad (TDAH) afecta comúnmente a niños de hasta 12 años, quienes pueden llegar a padecer uno de los tres subtipos de dicha patología; hiperactivo/impulsivo, inatento y combinado. La persona capacitada para diagnosticar esta patología es el psicólogo experto en el área, ya que se requiere evaluar a los niños, padres y profesores usando test estandarizados. Para que el especialista precise una conclusión debe realizar un cálculo de sumatoria y comparar el puntaje en un barómetro, que señala el rango de cada una de las subclases.

La clasificación de estos subtipos es una dificultad que las técnicas de Machine Learning (ML) pueden ayudar a solucionar, esto les permite a los especialistas tener un soporte para obtener resultados rápidos, que para problemas como estos, es una variable importante. Diagnosticar y tratar a los niños a tiempo evita que esta enfermedad persista hasta la edad adulta. Con los algoritmos de ML se puede llegar a identificar a los menores que requieran ayuda de terapia inmediata, y progresivamente puedan mejorar su rendimiento escolar, lo cual es una de las principales problemáticas que preocupan al Estado.

1.1 Descripción del problema

El problema del TDAH incluye una variedad de problemas asiduos, tales como la falta de concentración, hiperactividad y comportamientos impulsivos.

Este trastorno ha tenido una mayor tendencia en los últimos años, sin embargo, eso no implica que en el pasado este no haya existido, pues antes a los niños que probablemente padecían dicha enfermedad o presentaban los síntomas, eran calificados como “malcriados” o “vagos”, y en ciertos casos, los padres repercutían al castigo físico para corregir este comportamiento observando a menudo que el niño no

cambiaba su manera de actuar.

Aunque no haya cura para esta enfermedad, un diagnóstico y tratamiento temprano puede ser de gran importancia para inhibir su continuidad en el infante. Hay que tener claro que en muchos de los casos, padres y maestros no están orientados o desconocen de esta enfermedad por lo que podrían rotular erróneamente a un niño de inquieto, vago o hiperactivo, pero aquello no es correcto sin un diagnóstico. Fioravante et al. (2022)

Actualmente, en el país no se cuenta con las herramientas suficientes para ayudar a los niños con TDAH, pues un buen diagnóstico requiere test psiquiátricos que usan barómetros para decidir que subclase padece el paciente. Pero realizar esto a una gran cantidad de niños requiere mucho tiempo, y esta variable es un factor importante para realizar un tratamiento temprano. Si de clasificación se trata, los algoritmos del ML proveen soluciones óptimas capaces de dar soporte a los psicólogos o especialistas del tema.

Al igual que otro problema social en el país, la falta de información o investigación es un factor importante. Se recuerda que los actores involucrados en la problemática (sin incluir al psicólogo experto) nunca recibieron un asesoramiento para reconocer y ayudar en este tipo de trastornos. En Ecuador no existen muchos especialistas dedicados a explorar el TDAH, sus síntomas, diagnósticos y terapias, de modo que llegar a todas las escuelas a nivel nacional se ve como un sueño o meta poco alcanzable. La falta de compromiso social retarda la solución al problema en general.

Para entrenar un algoritmo de ML se requiere de la recopilación de datos clínicos, sin embargo, se dispone de un conjunto de datos de escuelas particulares y nacionales de la ciudad de Cuenca que será de utilidad para este trabajo, se debe depurar para su procesamiento.

1.2 Justificación del problema

En el Ecuador, en el campo de la psicología el TDAH ha sido poco explorado. Los niños que padecen del trastorno siguen siendo rotulados de “malcriados” y peor aún, no tienen una terapia adecuada. La desinformación conduce a un mal diagnóstico. Si se lograra entrenar un algoritmo que pueda clasificar de manera adecuada los resultados del test de

los niños, ayudaría a los psicólogos como un soporte para obtener los diagnósticos más rápido, permitiendo una investigación más eficiente para la obtención de tratamientos especializados para los niños.

Una rápida generación de datos permitirá, mediante técnicas estadísticas encontrar patrones y correlaciones que sean de ayuda para los especialistas.

Esto ayudará a que los niños que padecen de este trastorno puedan recibir una terapia personalizada acorde a las cuatro clasificaciones que ofrece el test de Conners, con ello los infantes podrán reintegrarse a la educación y mejorar su rendimiento académico. Se debe tomar en cuenta que el aumento de la demanda de evaluación aventaja a la cantidad de expertos médicos, esto implicaría que los pacientes no están siendo tratados rápidamente. Lograr que todos los psicólogos del país pueden realizar un diagnóstico rápido y eficiente tal vez no solucione el problema, pero si lo reduce.

1.3 Objetivos

1.3.1 Objetivo General

Adaptar un algoritmo que clasifique de manera óptima a niños de primaria según la categoría de trastorno de déficit de atención e hiperactividad mediante técnicas de Machine Learning.

1.3.2 Objetivos Específicos

A continuación se detallan los diferentes objetivos específicos del trabajo.

- Seleccionar las variables de mayor referencia para la implementación del algoritmo de clasificación tomando en consideración la opinión de los expertos;
- Decidir un algoritmo adecuado analizando su matriz de confusión.;
- Implementar el algoritmo para su fácil manejo por parte de los especialistas en la determinada área.

1.4 Marco teórico

El TDAH es un trastorno poco tratado en el Ecuador por los psicólogos, lo desconocen aún más los actores involucrados como lo son, los padres y profesores, pues no existen estudios que demuestren que los maestros pueden identificar con precisión a los niños que padecen el trastorno.

La detección de TDAH en un menor se lo realiza comúnmente mediante cuestionarios de diversas índoles y también notas clínicas como en Chen et al. (2021), otros investigadores prefieren encontrar una anomalía en el proceder usual del cerebro mediante radiografías o neuroimágenes. Pero gran parte de las investigaciones tiene como objetivo determinar el diagnóstico de un paciente haciendo uso de la electrofisiología y algoritmos de ML, para clasificar a los niños sin la necesidad de otras fuentes. Saha and Sarkar (2022)

Obtener datos electrofiológicos de niños en el Ecuador no es sustentable, debido a los recursos requeridos para su uso, por lo que los cuestionarios estandarizados, las entrevistas, las observaciones son las opciones más destacadas y necesarias, y fácilmente disponibles, de hecho el algoritmo de ML “árbol de decisión” presentó una efectividad de aproximadamente 85% con los datos estructurados como se mencionó anteriormente. Chen et al. (2021)

Chen, T., Antoniou, G., Adamou, M., Tachmazidis, I., & Su, P. (2021) tenían una colección amplia de datos que incluía siete cuestionarios (entre ellos las escalas de clasificación de Conner), y notas clínicas que podían estar estructuradas o no estructuradas. El objetivo de ellos era obtener el algoritmo de ML con mejor rendimiento, para seleccionarlo se basaron en la precisión (porcentaje de las predicciones correctas) y también tomaron en cuenta “el área bajo la curva característica operativa del receptor (AUC)”, con seis modelos de aprendizaje. Chen et al. (2021)

Se dio a notar que a veces incrementar el número de variables no suele mejorar el rendimiento de los modelos de aprendizaje, pues cuando se añadieron las variables de las notas médicas con un previo análisis de texto, estos modelos no se vieron beneficiados. Es necesario decidir manualmente que variables son las que dan una mayor calidad al algoritmo, cuando se añadieron las variables de las notas médicas con un previo análisis de texto estos modelos no se vieron beneficiados. Chen et al. (2021)

El presente estudio se realiza sobre los niños de segundo, cuarto y sexto año de educación básica, quienes en conjunto con sus padres y maestros realizaron un test estandarizado, con sus respectivos barómetros, para determinar si el niño padece del trastorno. Los datos recolectados presentan dos tipos de test para el TDAH, los cuales muestran resultados específicos distintos, pues existen dos condiciones heterogéneas del trastorno, el déficit de atención y la hiperactividad o impulsividad.

Al final de cada prueba realizada al niño, y a sus padres y profesores, mediante un barómetro, el psicólogo se encarga de ver el diagnóstico final. Sin embargo, sería de gran ayuda y un inmenso soporte que los ordenadores sean capaces de clasificar a los pacientes una vez terminada la prueba. Un algoritmo de clasificación de este tipo permitirá evitar a que existan fallas, no solo a la hora de diagnosticar sino también en errores de digitación, con ello se puede evitar dar falsos positivos o falsos negativos.

Con esta investigación, se busca adaptar un modelo de clasificación para diagnosticar a los pacientes con base en los test que se realicen, incluso se pretende hacer varios modelos para seleccionar el de mejor rendimiento para lograr una mayor eficiencia, también se busca observar el modelo que funciona óptimamente para cada test.

Los algoritmos de machine learning a usarse son K-Means Clustering, K-Nearest Neighbors Algorithm y Support Vector Machine Algorithm, cada uno de ellos clasifica datos de manera diferente, es por ello que se menciona anteriormente que se requiere indagar en la lectura para conocer más acerca de los tests de Conners y de Brief y dialogar con expertos que sugieran las variables de mayor influencia a la hora de dar un diagnóstico final.

Los algoritmos de ML permitirán diagnosticar a los pacientes escolares sin la necesidad de usar los barómetros de los test de TDAH, siendo un soporte para los psicólogos. Específicamente, automatizar la clasificación para diferenciar principalmente entre el TDAH y los niños que tienen un desarrollo típico.

El Support Vector Machine, usando datos clínicos de los registros médicos que incluyen además de datos demográficos, clasificaciones de síntomas de padres y maestros de Conners, entre otros, muestra una precisión que no puede ser consideraba como aleatoria. De acuerdo a una combinación de características de los pacientes se podría

mejorar el rendimiento del modelo de aprendizaje, es decir, tratar individualmente cada una y mediante una métrica cuantificar su importancia. Mikolas et al. (2022)
La colección de datos está diseñada para distinguir entre pacientes con TDAH y los que no lo padecen, es decir, se tiene como objetivo clasificar a los menores que poseen un posible trastorno, pues "los estudios previos que tenían como objetivo distinguir más de un trastorno de los controles de desarrollo típico informaron precisiones de clasificación más bajas que los estudios que tenían como objetivo clasificar a los individuos con desarrollo típico y a los pacientes con una afección". Mikolas et al. (2022)

1.5 Definiciones claves

Según el diccionario de la Real Academia Española, se entiende por desatender, a, "no prestar atención a lo que se dice o se hace" o "no hacer caso o aprecio de alguien o de algo". El déficit de atención es la escasez de la acción de atender, es decir, un niño de esta índole en muchas ocasiones tiene las características mencionadas. Real (2001)
Por otro lado, la Hiperactividad definida por la Real Academia Española es la "conducta caracterizada por un exceso de actividad". Nótese que las definiciones dadas solo pueden ser cuantificadas mediante una medición estándar, es por ello que se presenta la de definición de Barómetro: "Cosa que se considera índice o medida de un determinado proceso o estado". Real (2001)

1.6 Tests para diagnosticar TDAH

El test de Connors posee cuatro tipos de diagnósticos para los infantes: desarrollo típico, combinado, inatento e hiperactivo. Con 62 variables en total que brindan información de los niños con respecto a ciertas evaluaciones, además, sus padres y maestros aportan con datos de los menores desde su punto de vista para dar una clasificación.
Entre las variables más importantes destacamos: el nivel educativo, que nos brinda información sobre el año en curso de los pacientes; sexo, aciertos en exactitud de letras, su tiempo en exactitud de letras, aciertos en exactitud de palabras y tiempo de velocidad de palabras, además, dislexia visual, que es una variable categórica.
También tenemos el número de aciertos de exactitud de pseudopalabras, su

correspondiente tiempo y la variable dislexia fonológica. Otras cuatro variables más como: ojo catalejo, mano de escritura, notas y coeficiente intelectual completan las variables importantes en el test de Conners para los niños.

Las variables importantes en el Test aplicado a los padres ofrecen información desde otro punto de vista de los niños sobre problemas de aprendizaje, inatención, hiperactividad, funciones ejecutivas, agresividad, relación con los padres, impresiones positivas y negativas. Estas son más importantes porque ya están a escala.

Las variables obtenidas desde el punto de vista de los profesores son las mismas que las de los padres, excepto que no se cuenta con funciones ejecutivas y en este caso se tiene la relación con el profesor y no con los padres. Entonces observamos que gran parte de los datos son cuantitativos que están a escala para que mediante un barómetro se de un diagnóstico final sobre el paciente, es necesario recalcar que dicho barómetro es diferente acorde a la edad y sexo del niño.

Con respecto al Test de Brief, se poseen más datos recolectados (89 columnas) que brindan información acerca de los pacientes a través de valores escalados. Estos se dividen en tres grupos, dado que cada uno refleja los resultados del cuestionario realizado al padre, al profesor y al niño, pero que en conjunto son usados para dar un diagnóstico para el menor.

De los padres y profesores se obtienen datos acerca de los niños sobre su trabajo de memoria, control de emociones, entre otras. En cambio, los niños son evaluados con pruebas asociadas a sus funciones ejecutivas como por ejemplo relaciones analógicas, cálculo numérico, razonamiento numérico, legibilidad de escritura, entre otros. De todas las 89 variables solo se usaron aquellas que están a escalas y nos brindan datos útiles para una correcta clasificación.

1.7 Algoritmos de Machine Learning para la clasificación

1.7.1 K- Medias de agrupamiento

La idea general es identificar grupos o conglomerados de puntos de datos en un espacio multidimensional. Bishop and Nasrabadi (2006)

Suponga que se tiene un conjunto de n observaciones $(x_1, x_2, \ldots, x_n)$ (en el contexto de

esta investigación, cada x_i representa un niño previamente evaluado con los cuestionarios), cada uno de estos puntos es de dimensión d, por lo que, $x_i = (x_{i1}, x_{i2}, .., x_{id})$, en la que cada x_{ij} es una variable que guarda una característica.
Este algoritmo trata de crear K $(k \leqslant n)$ grupos a partir de n observaciones dadas. Es decir, sea U el conjunto muestral $(N(U) = n)$, entonces:

$$\bigcup_{i=1}^{K} S_i \tag{1.1}$$

En donde, los conjuntos S_m son el resultado del algoritmo, además, cada uno de ellos posee un vector característico u_m conocido como centro o media del conjunto. De tal forma que el objetivo sea "encontrar una asignación de puntos de datos a grupos, así como los centros $\{u_k\}$, tal que la suma de los cuadrados de las distancias de cada punto a su vector más cercano u_k, sea la mínima".
Formalmente:

$$F(\boldsymbol{u}) = \min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - u_i\|^2 \tag{1.2}$$

Donde $u = (u_1, u_2, \ldots, u_k)$. Ahora, en la práctica computacional, que los puntos posean una dimensión muy grande complica la clasificación. Este modelo puede incluir un paso extra que nos sugiere el "K" ideal para realizar las agrupaciones, si no es necesario, lo que se hace es preparar los datos para introducirlos en el código. En este caso, se debe etiquetar a cada conjunto S_k para que en los resultados nos de la clasificación y no solo los K agrupamientos.

1.7.2 1.7.2 K- vecinos más próximos (K- nn)

Suponga que se tiene un conjunto de n observaciones $\{(x_1, t_1), (x_2, t_2), \ldots, (x_n, t_n)\}$ tal que $x_i \in \mathbb{R}^d$ y $t_n \in 1, 2, ..I$ la etiqueta asignada, I es el número de clases. Se describen de este modo, porque en este modelo se busca predecir la clase de un punto a partir de un aprendizaje que la máquina realizó con datos de entrenamiento.
K-nn necesita de una colección de datos ya clasificados para ajustarse y lograr predecir la etiqueta de t_n de una nueva observación. Tal data, se particiona en dos, una destinada para el entrenamiento y otra para la prueba del modelo.

En los datos de entrenamiento, cada par (x_m, t_m) tiene asignado un t_m, mientras que en los datos de prueba se busca predecir qué valor tomará t_m, y por ende, a qué grupo pertenece x_m.

Entonces, se quiere determinar a qué clase pertenece una muestra de los datos de prueba x_j, para ello se escoge los k vecinos más cercanos y se le asigna la etiqueta $t_j = r$ a x_j si el grupo $(S_n, n \in 1, 2, .., I)$ con más representantes de los k vecinos es S_r.

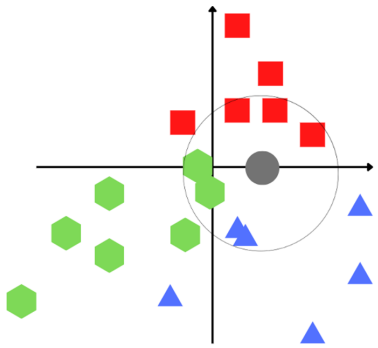


Figura 1.1. Se busca asignar una etiqueta a la muestra (círculo gris), en este caso K=7, se toman los siete vecinos más cercanos, de los cuales el grupo de los cuadrados posee mayor número de representantes (3) por lo que la muestra se asigna a dicho grupo.

Para encontrar los vecinos más cercanos normalmente se usa la distancia euclidiana:

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1.3}$$

Sin embargo, es posible usar otras métricas como:

$$\text{distancia de Manhattan: } d(X, Y) = \sum_{i=1}^{n} |x_i - y_i| \tag{1.4}$$

$$\text{Norma P: } d_p(X, Y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \tag{1.5}$$

Para realizar este modelo en los lenguajes de programación como Python o R-studio, se necesita un conjunto de datos, para su implementación es recomendable normalizar la data. El valor de ”K” se lo puede decidir, pero dependiendo de este valor se obtienen agrupaciones distintas.

Los datos se dividen en datos de entrenamiento y de prueba, para hacer esto el

lenguaje de Programación R-studio ofrece un método aleatorio, logrando dividir la data de manera aleatoria, logrando así una mejor precisión.

1.7.3 Máquinas de Vectores de Soporte (SVM)

Suponga que se tiene un conjunto de puntos de entrenamiento $\{(x_1, t_1), (x_2, t_2), ...(x_n, t_n)\}$, en donde $t_n \in \{-1, 1\}$ indica la clase a la que pertenece el punto x_i. Un hiperplano se puede representar como el conjunto de puntos $x = (x_1, x_2, \ldots, x_n)$ que cumplen que:

$$\mathbf{w}^T\mathbf{x} - b = 0 \tag{1.6}$$

$$\text{Ó } w_1x_1 + w_2x_2 + ...w_dx_d - b = 0 \tag{1.7}$$

Donde, $w = (w_1, w_2, \ldots, w_d)$, los w_i son los valores que se buscan, y se pueden determinar minimizando una función de error de suma de cuadrados regularizada, la cual es:

$$J(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}\{\mathbf{w}^T\mathbf{x}_n - t_n\}^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \tag{1.8}$$

Donde $\lambda \geqslant 0$ es un parámetro de regularización. En ocasiones, el conjunto de datos no se puede separar mediante un hiperplano, por lo que, se usa el método del Kernel, que consiste en utilizar una transformación $\phi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^D \ (d < D)$ que toma los puntos x_i para que estén en un espacio de dimensión mayor $\phi(x_i)$ donde exista un hiperplano que puede separar a los conjuntos de puntos $\{\phi(x_i)\}$ en dos clases.

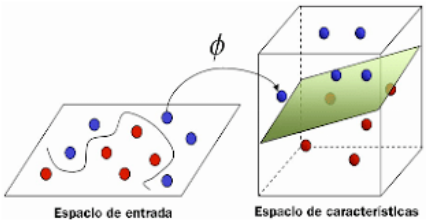


Figura 1.2. Interpretación gráfica del método del Kernel

En cuyo caso la función es:

$$J(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}\{\mathbf{w}^T\phi(\mathbf{x}_n) - t_n\}^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \tag{1.9}$$

Así, siempre hay un hiperplano que separa en dos partes el conjunto de datos. A pesar de que este método cumple espléndidamente su función, comete pequeños errores, que se buscan regular.

En la implementación del algoritmo, nuevamente se debe tomar en cuenta que las observaciones se dividen en datos de entrenamiento y en datos de prueba. Su funcionamiento es la de ajustar el hiperplano que logre la mejor separación y por lo tanto la mejor clasificación.

Los datos de entrenamiento son obtenidos de manera aleatoria de las n observaciones y se utilizan para buscar el hiperplano, este es el primer paso que se realiza en R-studio.

Posteriormente, con los datos de prueba se evalúa la efectividad del modelo o se calcula el porcentaje de error.

CAPÍTULO 2

2. METODOLOGÍA

En este capítulo se explica los pasos para encontrar una solución que permitió cumplir con los objetivos propuestos. Primero, se expuso el proceso de análisis de los datos, luego, se describió la investigación de los modelos de clasificación de Machine Learning. Posteriormente, se desarrolló la metodología con la que los algoritmos implementados en R-studio, clasificó en varios grupos (Subclases del TDAH) nuestros datos.

2.1 Descripción de la muestra

La base de datos posee 1535 observaciones, cada observación refleja resultados o características asociadas a un niño que fue evaluado con los cuestionarios de Conners y de Brief (También están incluidos las variables obtenidas por los padres y profesores del menor).

Todos estos datos los recolectó la Psicóloga Monserrate que reside en Cuenca . Varias instituciones educativas (Públicas y privados) de la ciudad fueron tomadas en cuenta para la investigación, los niños evaluados en ese entonces se encontraban cursando segundo, cuarto y sexto año de educación básica.

La información de los cuestionarios fue almacenada en un archivo Excel, que se dispuso. Como mencionamos en el anterior capitulo, gran parte de las variables eran de tipo cuantitativo.

2.2 Preparación de los datos

Los Tests usan barómetros para dar un diagnostico final, muchos cuestionarios psicológicos usan este modo, estos cambian sus índices acordes a la edad del niño. Es decir, un niño de seis años va a requerir un puntaje distinto al de un menor de ocho años para ser catalogado con desarrollo típico o normal.

En la literatura se observó que el género también influye en el diagnóstico final, razón

por la cual se decidió considerar este hecho para preparar los datos. Entonces, se tuvo dos condiciones que anunciaron que no se puede usar toda la data en conjunto para ser implementadas en los algoritmos de clasificación como *K-means, K-nn o SVM.*

En la charla con los expertos del área, se reconoció que los cuestionarios de Conners y de Brief presentaban información completamente distintas, así se reconoció tres razones por la cual dividir los datos. Se obtuvo los siguientes grupos: Segundo-hombres-Conners, Segundo-Mujeres-Conners, Cuarto-Hombres-Conners, Cuarto-Mujeres-Conners, Sexto-Hombres-Conners, Sexto-Mujeres-Conners, Segundo-hombres-Brief, Segundo-Mujeres-Brief, Cuarto-Hombres-Brief, Cuarto-Mujeres-Brief, Sexto-Hombres-Brief y Sexto-Mujeres-Brief, (Revisar la sección de abreviaturas en caso de ser necesario).

Se poseía entonces 12 conjuntos de observaciones, esto para que los modelos K-means, K-nn y SVM tuvieran mejores resultados. Cada uno de estos conjuntos se guardó en una hoja de Excel, ya que facilitó la implementación del algoritmo.

Luego, se procedió a solucionar el problema de los datos aberrantes o datos faltantes.

Se buscó datos atípicos como “999” o “99” que no tenían concordancia con los demás valores, en otros casos se observó que había datos faltantes (NA), la decisión dada fue de omitir aquellas observaciones cuyas variables tenían este problema.

2.3 Modelos de Machine Learning

De Bishop and Nasrabadi (2006), se aprendió que los modelos se ajustan con minimizar una función objetivo o función de error. También, que los modelos de K-means, K-nn y SVM funcionan con datos u observaciones que son similares a los datos que se tenían. Mikolas et al., 2022 menciona que el SVM es un algoritmo capaz de predecir el TDAH en un paciente, poseyendo información con respecto a cuestionarios de Conners y datos clínicos de los registros médicos.

Chen et al. (2021) propusieron que diagnosticar el TDAH es posible mediante árboles de decisión, pero para ello sus datos fueron distintos a los que se tuvieron en la investigación del presente documento.

2.4 Elección de variables

Se mencionó en el capítulo anterior, que cada niño presentaba más de 100 variables entre test de Conners y de Brief, pero no todas eran útiles para los algoritmos, además, un gran número de variables deduce una dificultad exponencial para clasificar observaciones usando algoritmos de clasificación como K-means, K-nn y SVM.

Los seis conjuntos de Conners (Segundo-hombres-Conners, Segundo-Mujeres-Conners, Cuarto-Hombres-Conners, Cuarto-Mujeres-Conners, Sexto-Hombres-Conners y Sexto-Mujeres-Conners) tenían 62 variables y los conjuntos de Brief, 89.

Se tuvo una charla con una experta en el tema en los cuestionarios de TDAH, además, se investigó conceptos asociados a la hiperactividad y déficit de atención para seleccionar las variables adecuadas.

El déficit de atención puede ser interpretado como la inatención, además, que esto se puede asociar con problemas de aprendizaje y falta de concentración. La hiperactividad tiene como foco principal el exceso de actividad, en el pasado los padres percibían este comportamiento como una acción mala, se indagó que una actitud agresiva tiene una relación con los comportamientos del niño.

Entonces se analizó todas las variables del Test de Conners, se eligió aquellas que están más relacionadas con lo mencionado en el párrafo anterior. También se decidió incluir las variables que reflejan el nivel de relación del niño con sus padres y profesores, pues esto se asocia al comportamiento del niño.

Por lo tanto, las variables seleccionadas para los primeros seis conjuntos fueron:

Variable	Significado	Rango (min - max)
$PD\_INATENCION\_PADRES$	Puntaje de inatención que los padres notan de su hijo	(0 - 15)
$PD\_HIPERACTIVIDAD\_PADRES$	Puntaje de Hiperactividad que los padres observan de su hijo	(0 - 18)
$PD\_LEARNING\_PROBLEMS\_PADRES$	Puntaje de problemas de aprendizaje que los padres observan de su hijo	(0 - 17)
$PD\_EJECUTIVE\_FUNCTIONS\_PADRES$	Puntaje de las funciones ejecutivas que observaron los padres	(0 - 15)
$PD\_AGRESIVIDAD\_PADRES$	Nivel de agresividad del menor percibida por los padres	(0-15)
$PD\_PEER\_RELATIONSHIPS\_PADRES$	Nivel de la relación que tienen los padres con el hijo	(0 - 18)
$PD\_POSITIVE\_IMPRESIN\_PADRES$	Grado de la impresión positiva que los padres tienen de sus hijos	(0-18)
$PD\_NEGATIVE\_IMPRESIOIN\_PADRES$	Grado de la impresión negativa que los padres tienen de sus hijos	(0 - 18)
$PD\_INATENTION\_TEACHERS$	Nivel de inatención que perciben los profesores de su estudiante	(0 - 16)
$PD\_HIPERACTIVIDAD\_TEACHERS$	Puntaje de hiperactividad que los profesores observa de su estudiante	(0-18)
$PD\_LEARNING\_PROBLEMS\_TEACHERS$	Puntaje de problemas de aprendizaje que los profesores observan de su estudiante	(0-18)
$PD\_AGRESIVIDAD\_TEACHERS$	Nivel de agresividad del menor percibida por los padres	(0-15)
$PD\_PEER\_RELATIONSHIP\_TEACHERS$	Puntaje de relación que los profesores tienen con su estudiante	(0-15)
$PD\_POSITIVE\_IMPRESIN\_TEACHERS$	Grado de la impresión positiva que los profesores tienen de su estudiante	(0-19)
$PD\_NEGATIVE\_IMPRESIN\_TEACHERS$	Grado de la impresión negativa que los profesores tienen de su estudiante	(0-18)

Tabla 2.1. Variables Test de Conners

Para el Test de Brief se necesitó más ayuda para la selección de variables. Con la ayuda de la experta, más la indagación de conceptos asociados a las variables del Test, se obtuvo un conjunto de variables importantes.

Finalmente se decidió que las variables a considerar se encuentran en una tabla C.1. en

anexos.

2.5 Implementación del Modelo K-Means

Este modelo toma una colección de observaciones y los clasifica en k grupos. Sin embargo, este valor k, dependió del Test que se empleó.

Para los seis data sets del Test de Brief, el valor de k fue dos, esto porque Brief solo clasifica al paciente con TDAH o desarrollo típico, por lo que el modelo tuvo que hacer la misma clasificación. Para el Test de Conners se tenía dos opciones para $k = 2, 4$, dado que el cuestionario ofrece cuatro tipos de diagnóstico.

Primero, se dividió la data en datos de entrenamiento y de prueba a través de una función de R-studio. Luego, los datos fueron normalizados con el objetivo de tener una mejor precisión. Lo que matemáticamente se hizo fue lo siguiente:

$$x_{escalado} = \frac{(x - \bar{x})}{S_x} \tag{2.1}$$

Siendo x el valor real de la observación, $\bar{x}$ es la media muestral y S_x la desviación estándar de la muestra. Con los datos escalados se averiguó mediante métodos como Elbow (método del codo), Silhouette y Gap Statistic, el valor óptimo de "k" para cada data set. Mount and Zumel (2019)

Para la aplicación de esos métodos, se necesitó de la matriz distancia descrita a continuación

$$\begin{pmatrix} 0 & d(x_2, x_1) & \cdots & d(x_n, x_1) \\ d(x_1, x_2) & 0 & \cdots & d(x_n, x_2) \\ \vdots & \vdots & \ddots & \vdots \\ d(x_1, x_n) & d(x_2, x_n) & \cdots & 0 \end{pmatrix} \tag{2.2}$$

Donde $d(x_i, x_j)$ es la distancia euclidiana de x_i a x_j. Como se mencionó anteriormente, el valor de k para Conners solo podía ser 2 y 4. El resultado de los tres métodos mencionados fue que el número de clúster debe ser $k = 2$.

Cuando $k = 2$, el algoritmo buscó los dos centros que realizaron la clasificación. Es decir, se encontró dos vectores representativos μ_1 y μ_2 para obtener los dos clústeres. Cuando

$k = 4$, el algoritmo encontró cuatro centroides.

Los pasos que realizó el algoritmo internamente fueron:

Paso 1: Cada punto es asignado al clúster cuyo centroide es el más cercano:

$$S_i^{(t)} = \{x : \|x - \mu_i^{(t)}\|^2 \leqslant \|x - \mu_j^{(t)}\|^2 \forall j \in [1, k]\} \tag{2.3}$$

Paso 2: Se actualizan los centroides de cada clúster para ajustar el modelo:

$$\mu_i^{(t+1)} = \frac{\sum_{x_j \in S_i^{(t)}} x_j}{\left|S_i^{(t)}\right|} \tag{2.4}$$

Deisenroth et al. (2020)

2.6 Implementación del Modelo SVM

Primero, se tomó de manera aleatoria el 70% de los datos para realizar el entrenamiento del modelo, el otro 30% se usó para probar el rendimiento del mismo.

De los 12 data sets, la mitad tenían cuatro categorías para clasificación (Hiperactividad, Inatención, Combinado y desarrollo típico) y la otra mitad dos categorías (Desarrollo típico y TDAH).

De esta manera, la función *factor()* en el código informa todas las categorías presentes en un Dataset. Por lo que, la estructura del seudocódigo no varía mucho entre todos los 12 conjuntos de datos.

R-studio tiene el modelo implementado en la función *tune()*. Luego de que se supo de las categorías se procedió a entrenar el modelo con dicha función. Mount and Zumel (2019) Finalmente, con el 30% de los datos de prueba se calculó los errores de entrenamiento, ya que los datos ya estaban categorizados para preparar el modelo. Se realizó más comandos para crear una matriz de confusión, pues se necesitó de esta herramienta para comparar modelos.

2.7 Implementación del Modelo K-NN

De forma aleatoria, se tomó el 70% de datos para entrenamiento y el otro 30% para la prueba del modelo. Resultó conveniente tener los datos de la forma (x_j, t_i) con t_i la categoría a la que pertenece x_j.

Con el código que se implementó en R-studio se obtuvo una clasificación y se realizó un gráfico de la misma. Pero, previamente se realizó la normalización de los datos.
Para el error, se utilizó los datos de prueba además de los de entrenamiento, pues la Dataset ya estaba categorizada. Se buscó crear la matriz de confusión para comparar el modelo en los 12 conjuntos de datos.

2.8 Análisis del porcentaje de error de cada modelo

Cada modelo de clasificación recibió 12 conjuntos de datos, y se obtuvo los resultados correspondientes, pero además se tuvo el porcentaje de error en cada uno, esto permitió entender que los resultados varían con la muestra.
La matriz de confusión se usó para realizar una comparación general entre los tres modelos de clasificación (K-means, K-nn, SVM). Pues el objetivo es saber cuál es el mejor con base en esta herramienta para evaluar el desempeño.

2.9 Gráficas del Modelo SVM

Realizar un gráfico de este modelo suele ser óptimo si se trabaja en un espacio de 2 o 3 dimensiones, pero los 12 conjuntos de datos tienen 15 dimensiones para los Dataset de Conners y 31 dimensiones para los restantes.
La alternativa para graficar el trabajo hecho por el modelo fue, tomar 2 variables al hacer y realizar una proyección del hiperplano con base en esas dos variables para tener una gráfica en 2 dimensiones del resultado.

2.10 Matriz de Confusión

La matriz de confusión es una herramienta que fue útil para visualizar el desempeño de los modelos de clasificación implementados Mount and Zumel (2019). Generalmente, permitió ver los aciertos y desaciertos de cada algoritmo Para el cuestionario de Conners, se implementó una función de R-studio (la cual recibió de entrada la información de las categorías reales y las predichas por el modelo) que devolvió como resultado una matriz de las características de la figura 2.1.

		Valor predicho			
Valor Real		Desarrollo típico	Inatención	Hiperactividad	Combinado
	Desarrollo típico				
	Inatención				
	Hiperactividad				
	Combinado				

Figura 2.1. Matriz de Confusión para los conjuntos de datos con el cuestionario de Conners

Para el caso de los grupos con el Test de Brief se tuvo una matriz de confusión de dimensiones más pequeñas.

		Valor Predicho	
		TDAH	Desarrollo típico
Valor Real	TDAH		
	Desarrollo típico		

Figura 2.2. Matriz de Confusión para los conjuntos de datos con el cuestionario de Brief

CAPÍTULO 3

3. RESULTADOS Y ANÁLISIS

3.1 Análisis de resultados

En este capítulo se detallan los resultados obtenidos en el presente proyecto integrador. Antes de dar comienzo al análisis de los resultados, se explica el tratamiento de los datos que se llevó a cabo pprevio a realizar la implementación del modelo matemático.

En los datos proporcionados por la Psi. Ximena Monserrat Vélez constaba la información de 1535 pacientes escolares pertenecientes a: segundo, cuarto, y sexto año de educación básica; de escuelas privadas y públicas de la ciudad de Cuenca. Dichos pacientes fueron sometidos a dos pruebas para diagnosticar TDAH, las cuales fueron: prueba de Conners, y prueba de Brief. Ahora bien, se procedió a dividir los datos de los pacientes en 12 grupos, porque los indicadores de las variables que se utilizan para diagnosticar el TDAH varían de acuerdo con la edad, el género, resultando así los 12 grupos.

3.2 Máquina de Vectores de Soporte

El modelo matemático SVM utiliza vectores de soporte como referencia a un subconjunto de observaciones de entrenamiento que identifican la ubicación del hiperplano de separación. A continuación, detallaremos la cantidad de vectores soportes que utilizó el algoritmo matemático en cada grupo, en particular para la prueba de Conners para hacer la clasificación.

Vectores de Soporte / Grupo	Combinado	Hiperactivo	Inatento	Desarrollo típico	Total
SHC	8	15	3	52	78
SMC	17	8	25	56	106
CHC	12	9	16	57	94
CMC	5	15	35	68	123
XHC	8	24	4	68	104
XMC	18	18	8	45	89

Tabla 3.1. Vectores Soportes pertenecientes a la prueba de Conners.

Se puede identificar que el grupo con mayor cantidad de vectores de soporte es el grupo de CMC, con un total de 123, lo que implica que este grupo presentó una mayor complejidad al momento de clasificar, esto se vio reflejado en el porcentaje de acierto del algoritmo. También se observó que el grupo de SHC es el que menor vectores de soporte necesita. Ahora, se detallan los vectores soporte que se obtuvieron al momento de clasificar a los pacientes usando la prueba de Brief.

Vectores Soporte / Grupos	Tiene TDAH	No tiene TDAH	Total
SHB	25	64	89
SMB	33	69	102
CHB	36	78	114
CMB	47	103	150
XHB	31	58	89
XMB	31	54	85

Tabla 3.2. Vectores Soportes pertenecientes a la prueba de Brief.

En esta prueba se identifica que el grupo que tiene mayor número de vectores de soporte es el grupo CMB con un total de 150. Al igual que en la prueba anterior, vemos que corresponde al mismo grupo que con Conners. Esto da una referencia de que el algoritmo, al momento de realizar la clasificación de este grupo, presenta una mayor complejidad, debido a que los rangos establecidos en cada variable para el grupo en cuestión no están bien definidos. Por último, se visualizó que el grupo que presenta

menor número de vectores de soporte es el grupo SMB con un total de 85.

3.3 Porcentaje de eficacia

El porcentaje de eficacia en un algoritmo es el indicativo de qué tan bien trabaja el modelo para solucionar el problema en cuestión. Se calculó el porcentaje de eficacia de cada uno de los 12 grupos, tanto para la data de entrenamiento la cual se trabajó con el 80% de los datos, y para la data de prueba la cual se trabajó con el resto de los datos. La siguiente tabla detalla los resultados obtenidos.

Porcentajes / Grupos	Grupo de Entrenamiento (%)	Grupo de Prueba (%)
SHC	0	7.5
SMC	2.38	10
CHC	2.79	11.32
CMC	2.47	10.34
XHC	1.37	9.61
XMC	0.44	16.66
SHB	0.62	5.12
SMB	2.72	11.42
CHB	6.1	13.46
CMB	1.69	25.42
XHB	0.55	6.66
XMB	7.81	15.55

Tabla 3.3. Porcentajes de error de los 12 grupos de estudio tanto para la data de entrenamiento como para la data de prueba.

Como observamos, el grupo que presenta mayor error es el de Cuarto Mujeres Brief con un 25.42%, seguido del grupo de Sexto Mujeres Conners con un 16.66%, y cerrando el tercer puesto el grupo se Sexto Mujeres Brief con un 15.55%. Estos resultados nos demuestran que el modelo tiende a fallar cuando trata de clasificar a las mujeres. Otro

conclusión que se puede obtener es que los porcentajes de error con los datos de entrenamiento son bajos lo cual son aceptables para nuestra investigación.

3.4 Matriz de Confusión

La matriz mostró la cantidad de aciertos, los falsos positivos, y los falsos negativos, además, está estrechamente relacionada con el porcentaje de error, si el porcentaje de error es del 0% tendremos una matriz diagonal. En caso contrario, tendremos una matriz cuadrada no diagonal la cual nos indicará la cantidad de falsos negativos, y la cantidad de falsos positivos. A continuación, se detalla la matriz de confusión de los 12 grupos de estudios, tanto para la data de entrenamiento, como para la data de prueba.

Caso Real / Predicción	Combinado	Hiperactivo	Inatento	Desarrollo típico
Combinado	8	0	0	0
Hiperactivo	0	16	0	0
Inatento	0	0	3	0
Desarrollo típico	0	0	0	139

Tabla 3.4. Matriz de confusión perteneciente al grupo de Segundo Hombres test de Conners, para los datos de entrenamiento.

Ahora, procederemos con la tabla de confusión de los datos de prueba.

Caso Real / Predicción	Combinado	Hiperactivo	Inatento	Desarrollo típico
Combinado	1	0	0	0
Hiperactivo	0	2	0	0
Inatento	0	0	0	0
Desarrollo típico	1	2	0	34

Tabla 3.5. Matriz de confusión perteneciente al grupo de Segundo Hombres test de Conners, para los datos de Prueba.

En esta tabla de confusión se ve que no hay una matriz diagonal, ya que un caso lo clasificó como Desarrollo Típico cuando el caso real es que ese paciente pertenece al grupo Combinado, mientras que existe cuatro pacientes que pertenece al grupo Hiperactivo. Sin embargo, el algoritmo arrojó dos falsos negativos como desarrollo típico. Las demás tablas de confusión se agregarán en los anexos del documento.

3.5 Gráficos de modelo SVM

A continuación, ilustra con un gráfico como actúa el modelo matemático, al momento de clasificar las observaciones, para este caso se tomó dos variables aleatoriamente para realizar el gráfico, ya que los programas no pueden graficar en más de tres dimensiones.

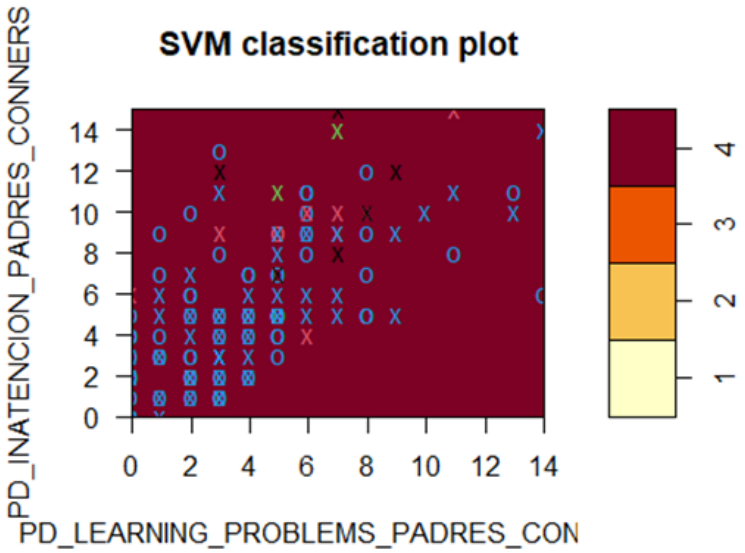


Figura 3.1. Gráfico del modelo SVM grupo de Segundo Hombres test de Conners

Este gráfico brinda una idea de que función del método del kernel utilizar para realizar el modelo, ya sea linear, radial, o polynomial. Los demás gráficos corresponde a las figuras B.1, B.2, B.3, B.4, B.5, B.6, B.7, B.8, B.9, B.10, B.11 que se los colocarán en la sección de anexos para su debida interpretación.

CAPÍTULO 4

4. CONCLUSIONES Y RECOMENDACIONES

En esta seción, se exponen las conclusiones y recomendaciones obtenidos luego del desarrollo de este trabajo. Se utilizó tres modelos de ML, los cuales permitieron un gran avance en la primera fase del problema del TDAH, el proceso de diagnosticar. Cuya dificultad es una consecuencia de la falta de recursos humanos, económicos y de tiempo.

Conclusiones

El motivo de este trabajo fue construir y entrenar modelos de clasificación para diagnosticar pacientes con TDAH utilizando como base de datos los resultados obtenidos de los encuestados mediante los Tests de Conners y de Brief. Se desarrollo con el objetivo de que el porcentaje de predicción sea considerablemente bueno (>85%) para que los expertos tengan la seguridad de implementar estos algoritmos en sus procesos de diagnósticos.

Con base en los resultados descritos en el capítulo tres se evidencia que se logró obtener el algoritmo que mejor clasifica a los niños de primaria, esto permite que los psicólogos tengan un soporte en este modelo matemático para diagnosticar más rápido a los pacientes escolares, sin la necesidad de acudir a barometros especializados, los cuales varían según la edad y género.

Los modelos K-medias, K vecinos más cercanos y Máquinas de vectores de soporte tienen sus respectivas ventajas y desventajas, con los datos con los que se trabajó se puede percatar en la matriz de confusión que el SVM ofrece mejor rendimiento. Además, investigaciones similares usan este algoritmo para predicciones. Por lo que, diagnósticar a los pacientes escolares utilizando técnicas de ML es posible.

El modelo SVM se entrenó con 12 conjuntos, se observó que es mucho más difícil reconocer si una niña padece de TDAH o no, pues la matriz de confusión no tenía la característica de ser diagonal, como en todos los algoritmos de ML, se ve necesario

recaudar más datos asociados a los pacientes femeninos para ajustar la precisión en la predicción de estos grupos.

Recomendaciones

- Aunque los resultados son bastante buenos, se debe tomar en cuenta que los modelos funcionan para la base de datos que se tiene, es decir, pueden existir conjuntos de datos atípicos que compliquen el funcionamiento y los resultados que se esperan obtener. Si se quiere que los algoritmos sean una herramienta fundamental, se debe recolectar una mayor cantidad de datos;
- El conjunto de datos que solo contenía los tests de las mujeres fueron más complicados de clasificar, por lo que se sugiere aplicar otros métodos de ML para estos casos, como lo son el árbol de decisión;
- Es posible que en el futuro se disponga de recursos económicos para obtener datos electrofisiológicos de aquellos pacientes que fueron diagnósticado con algún subtipo del TDAH, analizar la correlación y entrenar un algoritmo que identifique los casos con base en esos nuevos datos.

BIBLIOGRAFÍA

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Chen, T., Antoniou, G., Adamou, M., Tachmazidis, I., and Su, P. (2021). Automatic diagnosis of attention deficit hyperactivity disorder using machine learning. *Applied Artificial Intelligence*, 35(9):657–669.

Deisenroth, M. P., Faisal, A. A., and Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press.

Fioravante, I., Lozano, J. A. L., and Martella, D. (2022). Attention deficit hyperactivity disorder (adhd): A pilot study for symptom assessment and diagnosis in children in chile. *Frontiers in Psychology*, page 4772.

Mikolas, P., Vahid, A., Bernardoni, F., Süß, M., Martini, J., Beste, C., and Bluschke, A. (2022). Training a machine learning classifier to identify adhd based on real-world clinical data from medical records. *Scientific Reports*, 12(1):12934.

Mount, J. and Zumel, N. (2019). *Practical data science with R*. Simon and Schuster.

Real, A. E. (2001). Diccionario de la lengua española. *http://www. rae. es.*

Saha, P. and Sarkar, D. (2022). Characterization and classification of adhd subtypes: An approach based on the nodal distribution of eigenvector centrality and classification tree model. *Child Psychiatry & Human Development*, pages 1–13.

APÉNDICES

APÉNDICE A

A.1 Matriz de Confusión: Tests de Conners y Brief

Caso Real / Predicción	Combinado	Hiperactivo	Inatento	Desarrollo típico
Combinado	17	0	0	0
Hiperactivo	0	26	1	2
Inatento	0	0	6	0
Desarrollo típico	0	0	1	115

Tabla A.1. Matriz de confusión perteneciente al grupo de Segundo Mujeres test de Conners, para los datos de entrenamiento.

Caso Real / Predicción	Combinado	Hiperactivo	Inatento	Desarrollo típico
Combinado	3	0	0	0
Hiperactivo	1	5	1	0
Inatento	0	0	0	0
Desarrollo típico	0	1	1	28

Tabla A.2. Matriz de confusión perteneciente al grupo de Segundo Mujeres test de Conners, para los datos de Prueba.

Caso Real / Predicción	Combinado	Hiperactivo	Inatento	Desarrollo típico
Combinado	8	0	0	0
Hiperactivo	0	12	0	0
Inatento	0	1	11	0
Desarrollo típico	0	3	1	178

Tabla A.3. Matriz de confusión perteneciente al grupo de Cuarto Hombres test de Conners, para los datos de Entrenamiento.

Caso Real / Predicción	Combinado	Hiperactivo	Inatento	Desarrollo típico
Combinado	1	0	0	0
Hiperactivo	0	1	0	0
Inatento	1	0	1	0
Desarrollo típico	0	3	2	44

Tabla A.4. Matriz de confusión perteneciente al grupo de Cuarto Hombres test de Conners, para los datos de Prueba.

Caso Real / Predicción	Combinado	Hiperactivo	Inatento	Desarrollo típico
Combinado	14	0	0	0
Hiperactivo	0	32	0	0
Inatento	1	0	3	0
Desarrollo típico	1	3	2	188

Tabla A.5. Matriz de confusión perteneciente al grupo de Cuarto Mujeres test de Conners, para los datos de entrenamiento.

Caso Real / Predicción	Combinado	Hiperactivo	Inatento	Desarrollo típico
Combinado	2	0	0	0
Hiperactivo	0	4	0	0
Inatento	1	0	0	0
Desarrollo típico	0	4	1	46

Tabla A.6. Matriz de confusión perteneciente al grupo de Cuarto Mujeres test de Conners, para los datos de Prueba.

Caso Real / Predicción	Combinado	Hiperactivo	Inatento	Desarrollo típico
Combinado	8	0	0	0
Hiperactivo	0	22	0	1
Inatento	0	0	4	0
Desarrollo típico	1	2	0	182

Tabla A.7. Matriz de confusión perteneciente al grupo de Sexto Hombres test de Conners, para los datos de Entrenamiento.

Caso Real / Predicción	Combinado	Hiperactivo	Inatento	Desarrollo típico
Combinado	2	1	0	0
Hiperactivo	0	2	0	1
Inatento	0	0	0	0
Desarrollo típico	0	3	0	43

Tabla A.8. Matriz de confusión perteneciente al grupo de Sexto Hombres test de Conners, para los datos de Prueba.

Caso Real / Predicción	Combinado	Hiperactivo	Inatento	Desarrollo típico
Combinado	21	0	0	0
Hiperactivo	0	17	0	0
Inatento	0	0	9	0
Desarrollo típico	0	1	0	175

Tabla A.9. Matriz de confusión perteneciente al grupo de Sexto Mujeres test de Conners, para los datos de Entrenamiento.

Caso Real / Predicción	Combinado	Hiperactivo	Inatento	Desarrollo típico
Combinado	2	1	0	1
Hiperactivo	1	2	0	2
Inatento	1	0	1	0
Desarrollo típico	1	1	1	40

Tabla A.10. Matriz de confusión perteneciente al grupo de Sexto Mujeres test de Conners, para los datos de Prueba.

Caso Real / Predicción	TDAH	No TDAH
TDAH	24	0
No TDAH	1	134

Tabla A.11. Matriz de confusión perteneciente al grupo de Segundo Hombres test de Brief, para los datos de entrenamiento.

Caso Real / Predicción	TDAH	No TDAH
TDAH	4	0
No TDAH	2	33

Tabla A.12. Matriz de confusión perteneciente al grupo de Segundo Hombres test de Brief, para los datos de prueba.

Caso Real / Predicción	TDAH	No TDAH
TDAH	36	0
No TDAH	4	107

Tabla A.13. Matriz de confusión perteneciente al grupo de Segundo Mujeres test de Brief, para los datos de entrenamiento.

Caso Real / Predicción	TDAH	No TDAH
TDAH	6	1
No TDAH	3	25

Tabla A.14. Matriz de confusión perteneciente al grupo de Segundo Mujeres test de Brief, para los datos de prueba.

Caso Real / Predicción	TDAH	No TDAH
TDAH	26	2
No TDAH	11	174

Tabla A.15. Matriz de confusión perteneciente al grupo de Cuarto Hombres test de Brief, para los datos de entrenamiento.

Caso Real / Predicción	TDAH	No TDAH
TDAH	2	0
No TDAH	7	43

Tabla A.16. Matriz de confusión perteneciente al grupo de Cuarto Hombres test de Brief, para los datos de prueba.

Caso Real / Predicción	TDAH	No TDAH
TDAH	48	0
No TDAH	4	184

Tabla A.17. Matriz de confusión perteneciente al grupo de Cuarto Mujeres test de Brief, para los datos de entrenamiento.

Caso Real / Predicción	TDAH	No TDAH
TDAH	6	8
No TDAH	7	38

Tabla A.18. Matriz de confusión perteneciente al grupo de Cuarto Mujeres test de Brief, para los datos de prueba.

Caso Real / Predicción	TDAH	No TDAH
TDAH	31	0
No TDAH	1	151

Tabla A.19. Matriz de confusión perteneciente al grupo de Sexto Hombres test de Brief, para los datos de entrenamiento.

Caso Real / Predicción	TDAH	No TDAH
TDAH	5	0
No TDAH	3	37

Tabla A.20. Matriz de confusión perteneciente al grupo de Sexto Hombres test de Brief, para los datos de prueba.

Caso Real / Predicción	TDAH	No TDAH
TDAH	27	2
No TDAH	13	150

Tabla A.21. Matriz de confusión perteneciente al grupo de Sexto Mujeres test de Brief, para los datos de entrenamiento.

Caso Real / Predicción	TDAH	No TDAH
TDAH	2	1
No TDAH	6	36

Tabla A.22. Matriz de confusión perteneciente al grupo de Sexto Mujeres test de Brief, para los datos de prueba.

APÉNDICE B

B.1 Gráficos del modelo SVM

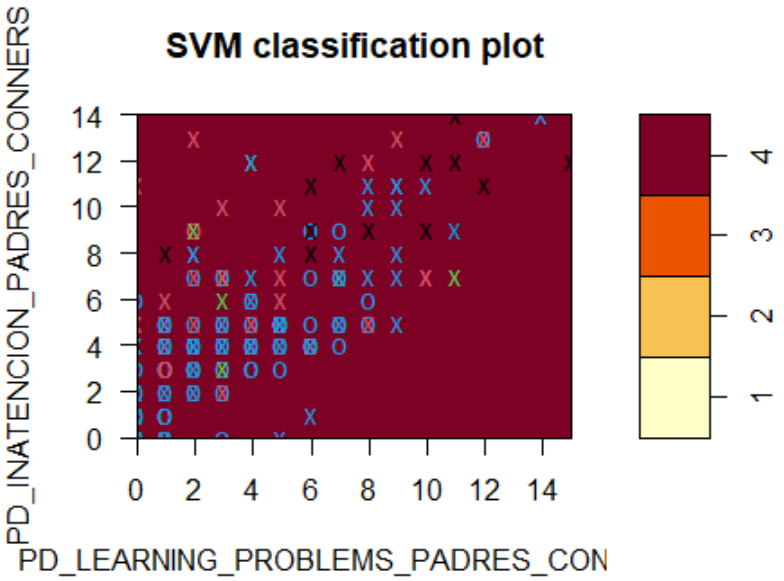


Figura B.1. Gráfico del modelo SVM grupo de Segundo Mujeres test de Conners.

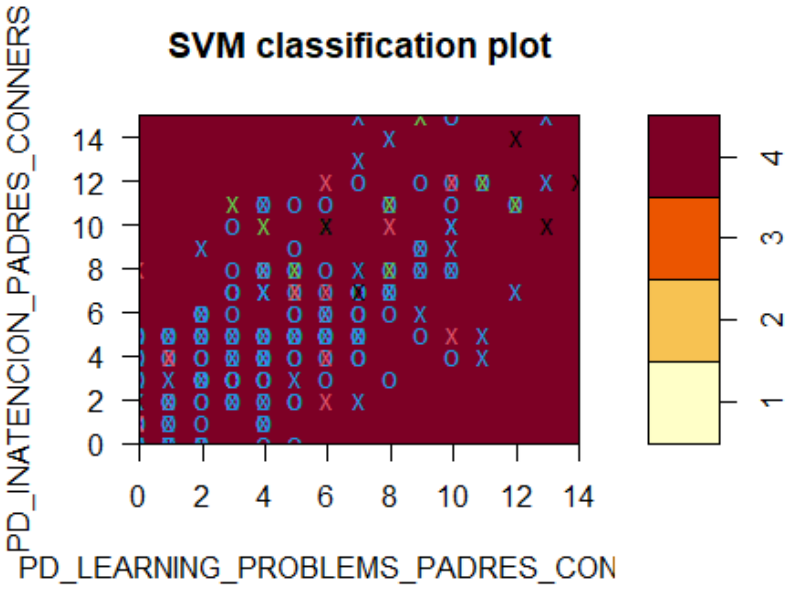


Figura B.2. Gráfico del modelo SVM grupo de Cuarto Hombres test de Conners.

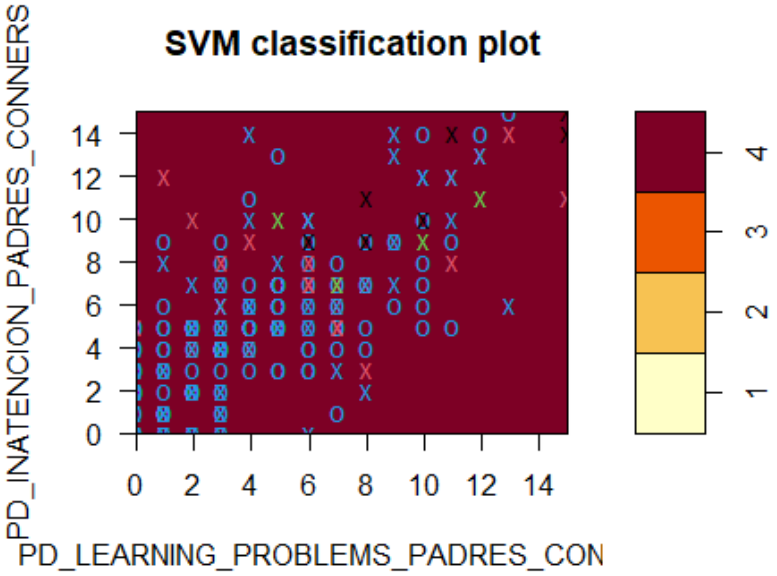


Figura B.3. Gráfico del modelo SVM grupo de sexto Hombres test de Conners.

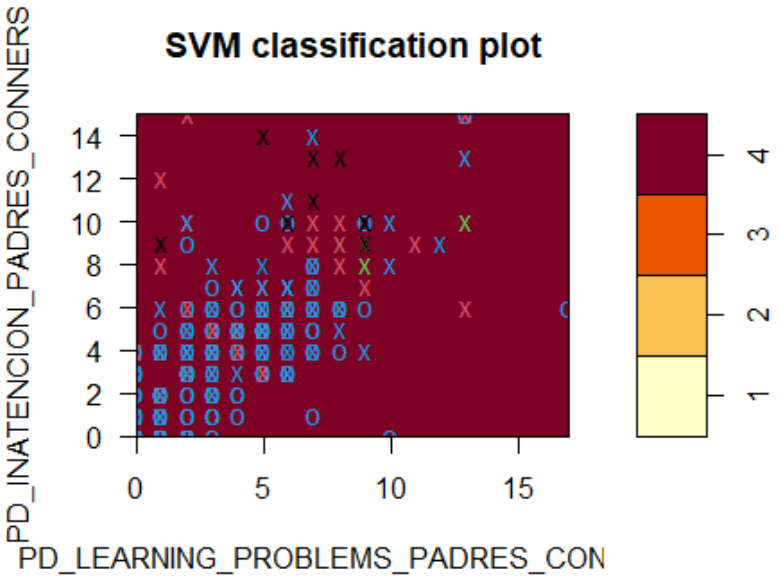


Figura B.4. Gráfico del modelo SVM grupo de cuarto Mujeres test de Conners.

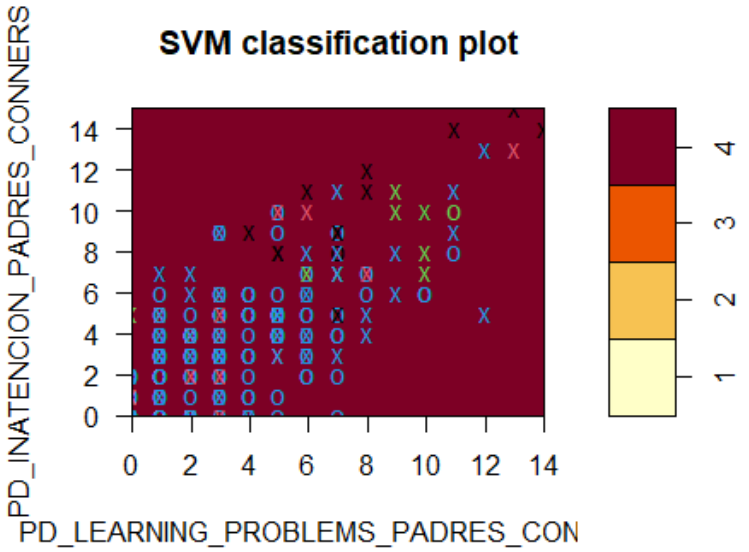


Figura B.5. Gráfico del modelo SVM grupo de sexto mujeres test de Conners.

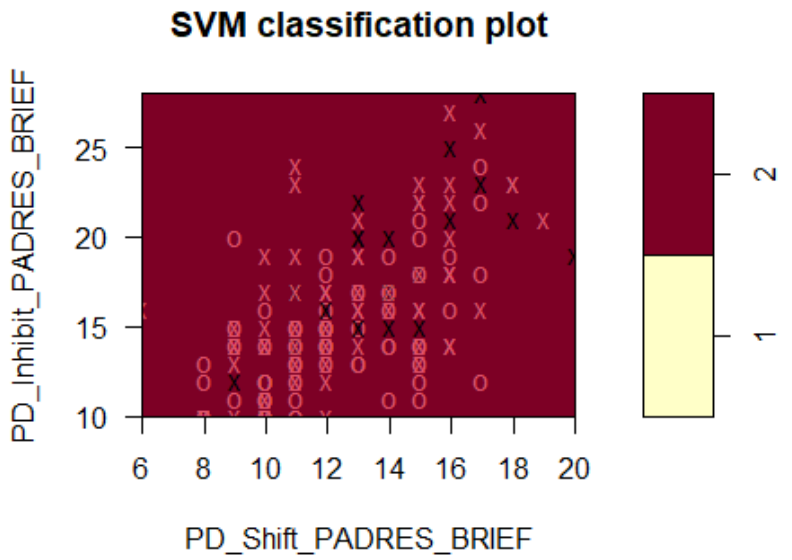


Figura B.6. Gráfico del modelo SVM grupo de Segundo Hombres test de Brief.

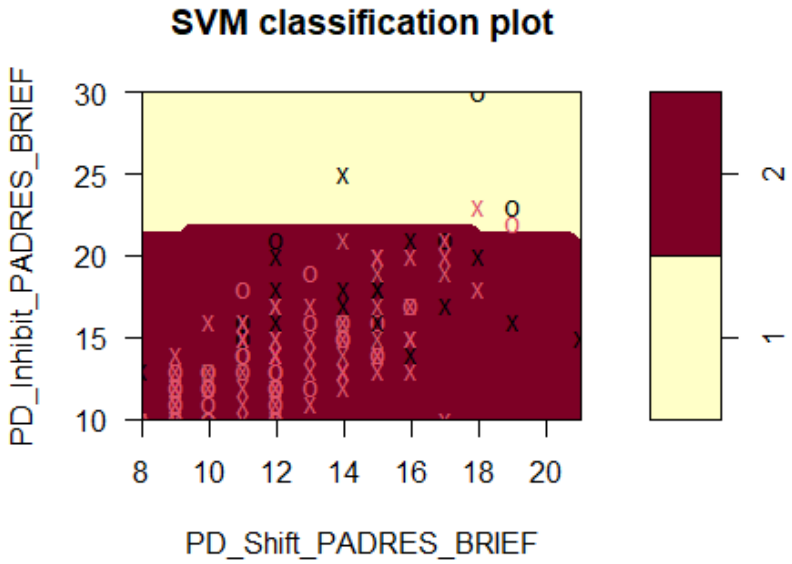


Figura B.7. Gráfico del modelo SVM grupo de Segundo Mujeres test de Brief.

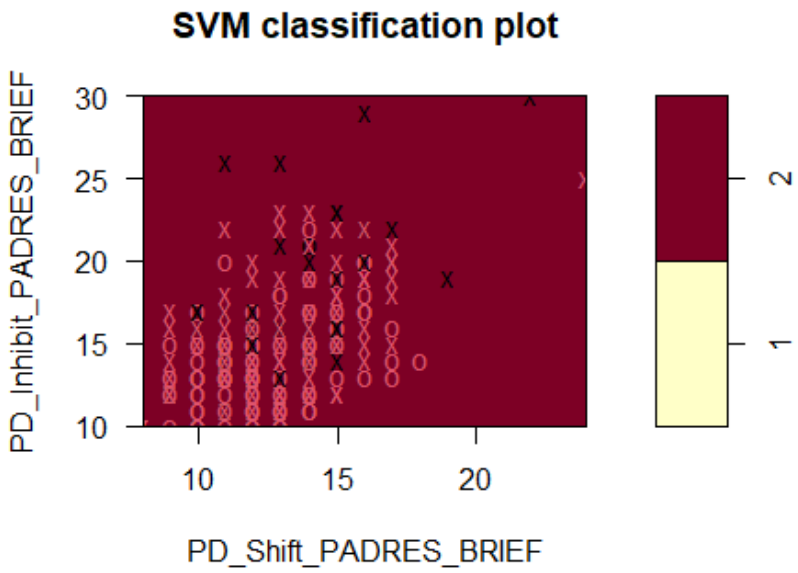


Figura B.8. Gráfico del modelo SVM grupo de cuarto hombres test de Brief.

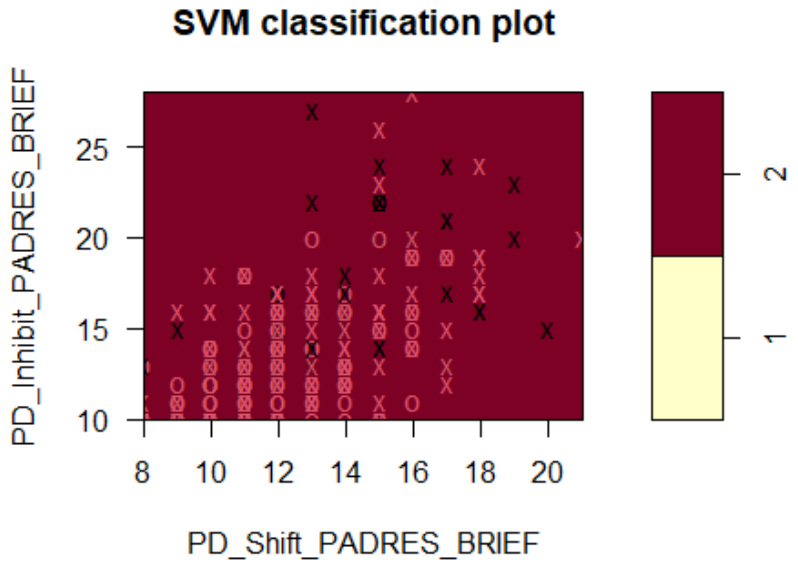


Figura B.9. Gráfico del modelo SVM grupo de Cuarto mujeres test de Brief.

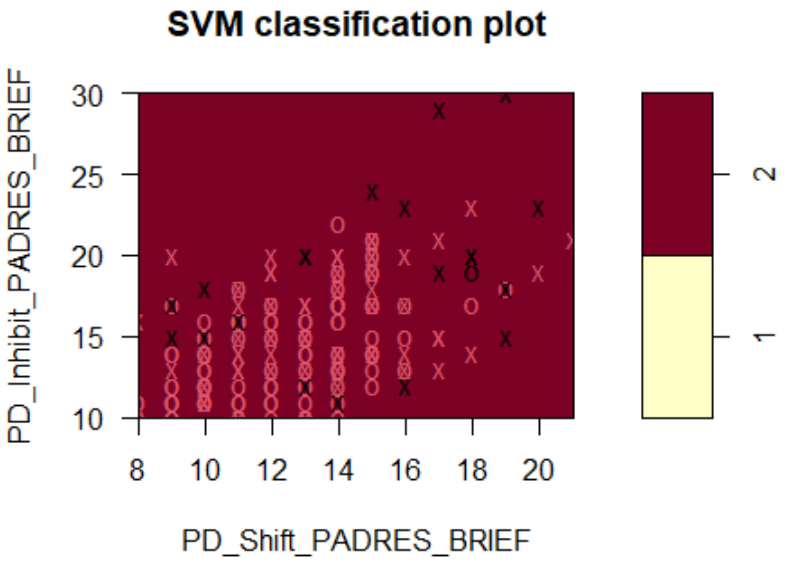


Figura B.10. Gráfico del modelo SVM grupo de Sexto Hombres test de Brief.

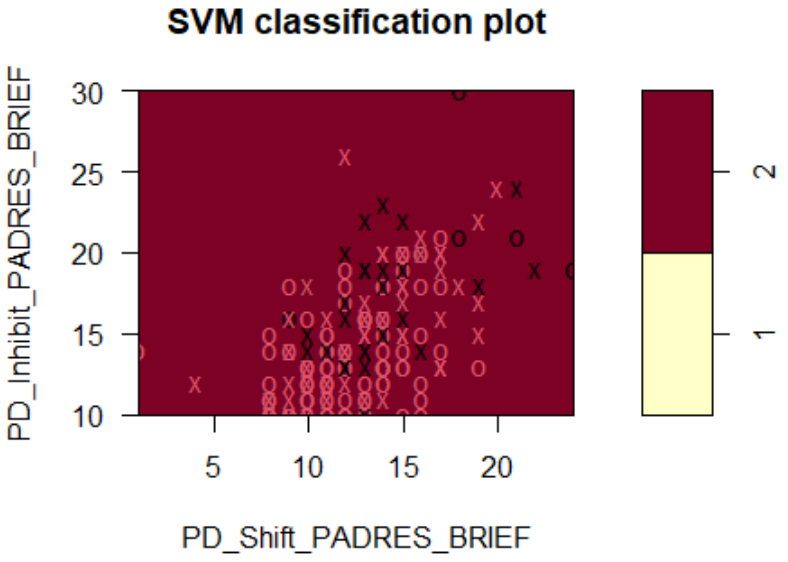


Figura B.11. Gráfico del modelo SVM grupo de sexto mujeres test de Brief.

APÉNDICE C

C.1 Variables del Test de Brief

Variable	Rango (min - max)
$PD\_INHIBIT\_PADRES$	(10 - 30)
PD_SHIFT_PADRES	(1-24)
$PD\_EMOTIONAL\_CONTROL\_PADRES$	(1 - 30)
PD_BRI_PADRES	(25 - 84)
$PD\_INITATE\_PADRES$	(7-24)
$PD\_WORKING\_MEMORY\_PADRES$	(4 - 30)
$PD\_PLAN\_ORGANIZE\_PADRES$	(10-36)
$PD\_ORGANIZATION\_MATERIALS\_PADRES$	(6 - 20)
$PD\_MONITOR\_PADRES$	(3 - 24)
PD_MI_PADRES	(8-169)
PD_GEC_PADRES	(41-216)
$PD\_INHIBIT\_TEACHERS$	(8-30)
$PD\_SHIFT\_TEACHERS$	(7-30)
$PD\_EMOTIONAL\_CONTROL\_TEACHERS$	(8-30)
$PD\_BRI\_TEACHERS$	(28-123)
$PD\_INITATE\_TEACHERS$	(5-21)
$PD\_WORKING\_MEMORY\_TEACHERS$	(3-35)
$PD\_PLAN\_ORGANIZE\_TEACHERS$	(1-30)
$PD\_ORGANIZATION\_MATERIALS\_TEACHERS$	(4-21)
$PD\_MONITOR\_TEACHERS$	(8-30)
$PD\_MI\_TEACHERS$	(37-132)
$PD\_GEC\_TEACHERS$	(65-219)

Tabla C.1. Variables Test de Brief

Variable	Significado	Rango (min - max)
PD_RELACIONES_ ANALOGICAS_BADYG	Puntaje de relaciones analógicas obtenido por el niño	(0 - 26)
PD_PROBLEMAS_ NUMERICOS_BADYG	Puntaje en problemas numéricos obtenido por el niño	(0 - 31)
PD_MATRICES _LOGICAS_BADYG	Puntaje en matrices lógicas obtenido por el niño	(0 - 29)
PD_ORDENES_VERBALES_ COMPLEJAS_BADYG	Puntaje obtenido por el niño en ordenes verbales complejas	(0 - 27)
PD_FIGURAS _GIRADAS_BADYG	Puntaje de identificación de figuras giradas, obtenidas por el niño	(0-27)
PD_CALCULO_ NUMERICO_BADYG	Puntaje de en aciertos de cálculo numérico obtenido por el niño	(0 - 24)
PD_MEMORIA_ RELATO_ORAL_BADYG	Puntaje en memoria de relato oral obtenido por el niño	(0 - 30)
PD_MEMORIA_V_ORT_ALTERACIONES_ _EN_LA_ESCRITURA_BADYG	Puntaje de memoria verbal y ortografía, y alteraciones en la escritura obtenido por el niño	(0 - 32)
PD_DISCRIMINACION_DE_ DIFERENCIAS_BADYG	Puntaje de capacidad para discriminar diferencias entre objetos	(0 - 88)
PD_INTELIGENCIA_ GENERAL_BADYG	Puntaje de inteligencia general del niño	(0 - 150)
PD_RAZONAMIENTO_ LOGICO_BADYG	Puntaje de la encuesta de razonamiento lógico del niño.	(0 - 91)

Tabla C.2. Variables Test de BRIEF